Summary and discussion

Saris, Willem E.; Kaase, Max

CHAPTER 11

SUMMARY AND DISCUSSION

WILLEM E. SARIS AND MAX KAASE

11.1 Introduction

In this book a series of studies has been reported related to the effects of different modes of survey data collection in the social sciences. The methodological study covered in this book was triggered by a decision of the European Commission in favour of a partial shift of the Eurobarometer from face to face data collection to telephone data collection. Since this switch can cause a variety of problems, the consequences of this change need to be explored. After the explanation of the design of the study, a detailed description of the data collection methods used and an overview of the different problems, in this chapter the results of the analysis will be summarised. After that, some practical conclusions will be drawn, and various scientific issues will be addressed in the concluding paragraphs.

11.2 The design of the methodological study

Given the relevance of the Eurobarometer data for all those interested in the development of political orientations in Europe and for social science research in general, the changes which were expected to occur in the Eurobarometer data because of the change in data collection modes was enough reason to suggest that methodological research should be done to evaluate the consequences of that change. The argument in favour of such research was based on prior knowledge on mode effects. For example, Groves (1989) gave nine reasons why one should expect differences between face to face and telephone data collection. They can be condensed as follows:

1. The coverage of the population will be different for face to face interviews and telephone interviews, since those people who do not have a telephone will not be representative of the general population from which the sample should be drawn. In Europe, this difference can be substantial because in some areas and countries household telephone density is close to 100% while in other areas and countries the coverage is closer to 50%. In the USA it has been found that this coverage error leads to considerable differences in responses on several dimensions (Groves and Kahn, 1979; Cannel et al., 1987).
2. The field work of the organisations doing the surveys can and usually will be quite different with respect to the interviewers used, their training and supervision, the number of times that a respondent is contacted, and the rules by which a refusal is accepted.

Differences in these management aspects will lead to differences in nonresponse and consequently to differences in findings.

3. The mode of data collection itself can also lead to different results. It is possible that people react differently to the same question in a telephone interview and in a face to face interview. For instance, it has been found that open-ended questions result in more elaborated answers in face to face interviews than in telephone interviews. Also, more acquiescence and an extremeness bias might be expected (Groves, 1989). However, the general picture is that these mode effects, after correcting for all other factors, are rather small (De Leeuw and van der Zouwen, 1988).

Also, mode-connected effects are possible, that is effects which might occur due to the fact that changes in the approach to the respondent are necessary depending on the mode of interview, and that these changes will matter. For example, the use of show cards is not possible in telephone interviews, and as a consequence the procedure for complex questions has to be adjusted. In order to cope with this problem, in telephone surveys, commonly a two-step approach is used where first a small number of crude categories is presented which are later split up into more differentiated ones. The idea here is to obtain the same kind of precision by telephone as in personal interviews where show cards are used to present ten or so categories at the same time. These mode-related changes in the questionnaire can lead to substantial discrepancies in the results, as has been shown by Groves and Kahn (1979), Miller (1984), and Monsees and Massey (1979).

This brief overview points to the main reasons why a change from face to face to telephone interviews will most likely lead to different results. As indicated in chapter 2, one can expect that the total difference (T) between face to face and telephone interview responses in percentages or in mean score will be equal to the difference due to coverage (C) plus the difference due to difference in nonresponse (N) plus the difference due to the mode of data collection (M):

$$T = C + N + M \quad (1)$$

Given the possible confusion due to discrepancies in results, notably the issue of errors in both procedures needs to be addressed more precisely, but also ways should be found to adjust the findings in such a way that the results become comparable.

In order to help with the methodological study, the Berlin-based FORSA research institute offered to collect data for a limited set of questions in all countries through telephone interviewing, while at about the same time the INRA institute conducted the standard Eurobarometer 41 face to face. This approach is a good simulation of the future situation in Europe when two studies will be done on identical topics at the same time by different survey organisations, each using a different data collection mode. The design used here permits to estimate the total difference in responses for two specific organisations (T).

There is, however, one major weak point here. The problem is that there are too many factors creating differences between the various approaches, and that one therefore cannot determine precisely which one causes these differences. Anticipating this lack of strength in the design, a panel element was included in this study.

Here, the respondents were first confronted in a face to face interview with the normal Eurobarometer questionnaire. In addition, they were asked whether they had a telephone and were willing to answer some questions some time later through the telephone. If they agreed, they were called back after about a week to respond to a small number of questions already put to them before in the Eurobarometer. This panel design offers better insights into the effects of the two different sources of error, as follows.

First, when the telephone owners and non-owners are compared, an estimate can be obtained of the effect of telephone ownership on the distribution of responses in the panel to the relevant variables. In this comparison no other variables intervene because the same people are studied and all questions are presented in a face to face interview. So the only possible explanation for differences is telephone ownership, and thus a good estimate of the coverage error (C) which will occur, is provided.

A second effect that can be studied with this design relates to the mode of data collection (M) since one can compare the answers of the respondents to the same questions in the personal interview and in the telephone interview. This evidence is not so strong as in the case of the comparison of telephone owners and non-owners because there are other factors besides the mode effects which can come into play.

This design does not allow for an independent estimate of the effects of the fieldwork organisation on the nonresponses (N), but one can at least deduce this effect. The direct comparison of personal interviews with telephone interviews gives an estimate of T. Using the panel design, C and M can be assessed. Using the combination of the two designs, the effect of the difference in nonresponse due to different organisational procedures will be:

$$N = T - C - M \qquad (2)$$

It should be recalled here that the coverage error (C) is an estimate which for the largest part is independent of the organisation that did the research because the effect is determined by the difference between owners and non-owners in the population. This difference will only minimally be influenced by the specific procedure used for data collection, as long as this procedure is not completely flawed.

The same point can be made for the estimate of the mode effect M as was argued above. On the other hand, the estimates for nonresponse (N) and for the total difference (T) are clearly influenced by the organisations which perform the studies. The total difference varies directly with the difference in nonresponse which is produced by the two organisations in question. So general statements are difficult to make about these two components although they can be properly assessed for a specific case.

Furthermore, the estimates of the coverage error and of the mode effect can also vary with the topics being addressed. Telephone non-owners can differ in their opinions on certain questions, and this will lead to differential effects although for other questions the differences can be very small. In the literature, some questions have been mentioned to be more effected than others, like open-ended questions, questions placing a heavy cognitive burden on respondents such as long questions or questions with a large number of categories. The same holds true for questions which are normally asked with a show card, a procedure presently not

available in telephone interviewing. Given the effects of the type of question asked, the different types were studied separately. For an overview of the questions we refer to the questionnaires in Chapter 1.

11.3 Results

In this section the results of this study will be summarised, following most of the time the sequence in which the results have been presented in the book.

11.3.1 Sample differences

In chapter 2 an overview has been given of the data collection procedures of the two survey organisations which collected the standard Eurobarometer data by face to face interviews and by telephone interviews. It has been shown that the procedures used were different on several points as can be seen in table 11.1.

Chapter 3 reported an effort to make the samples as comparable as possible by weighting on the basis of variables for which information about the population distributions is available. It has been shown that the weighting procedures could not reduce the serious differences which existed between the two studies on several variables. It seems that the correlation between the weighting variables, the variables which cause the problems and the variables of interest are not strong enough to produce an acceptable level of adjustment.

In chapter 4 the coverage error was studied in detail. It turned out that the group of respondents without a telephone can be very different from the group with a telephone. It must be understood, though, that this does not necessarily bias the results of telephone studies very much, especially when the size of the group is rather small.

In order to study this phenomenon, chapter 5 presented an effort to estimate the size of the different component of the total differences for the three participating countries. The result of this analysis has been reprinted in table 11.2. It should be noted that since the estimates of the

different effects are based on calculations over all categories of variables, equation 1 does not hold anymore (this equality holds for each category but not necessarily for the sum ignoring the signs).

When one looks at the size of the effects, a clear rank order can be established. Averaged across the three countries, the coverage differences rank lowest with a mean of 1.1 percentage points over all questions and countries. Mode differences are remarkably larger with an average score of 5.2 percentage points. However, it cannot be concluded that this is necessarily due to the telephone interviews. It may be that respondents produce more random answers in telephone interviews because the time pressure is stronger and they are not supported by visual aids. On the other hand, interviewers in telephone interviews are more controlled which means that they ask the questions more precisely in the way expected than in the uncontrolled face to face interview situation. Clearly, the largest differences come from

the black box of the fieldwork of the two survey organisations. The mean differences over topics and countries is 7.7. percentage points.

Table 11.1 Summary of the sampling methods

	EB41.0	**EB41.Panel**	**FORSA**
Type	face to face	telephone	telephone
Fieldwork	April 4th - May 6th	April 5th - April 30th	April 28th - June 3rd
Countries	12 EU member states	France, Belgium, Spain	12 EU member states
Completion rate	EU: 44,7%	37,6% (% of eligible households)	EU: 43,4%
Sample frame	- Census enumeration units (or otherwise)	- respondents of EB41.0 with a telephone who have given their number	Telephone directories
Selection Method	- more than 100 sampling units per country are randomly chosen as start address - a random increment provides up to 10 addresses - one person/per household selected by next birthday or Kish method or an other	- all possible respondents are contacted - controlled by Age, Sex, Occupation and Subjective Social Class	- From 10 to 22 'provinces' per country samples are drawn according to the size of the province's population - one person/per household selected by next birthday method
Interviewers' testing and supervision	INRA's national associates are responsible	INRA central - computerised dialling	FORSA central - tests in advance - computerised dialling
Call backs	2 revisits	8 call backs	12 call backs
Refusals	no refusal reversion	no refusal reversion	no refusal reversion
Substitution	random walk	no substitution	by random number

Table 11.2 A summary of all one-directional differences in three countries

		Total (T)	Coverage (C)	Mode (M)	Organisation (N)
Satisfaction with life	France	10.9	0.3	5.3	9.3
	Spain	21.9	0.9	5.0	21.7
	Belgium	4.6	2.0	16.6	12.7
Satisfaction with democracy	France	4.9	0.3	9.3	10.4
	Spain	8.8	1.7	5.4	9.4
	Belgium	3.4	0.8	10.8	13.9
Persuade others	France	17.1	0.6	3.6	15.8
	Spain	11.6	0.8	9.9	3.6
	Belgium	7.8	1.6	10.9	8.8
Political discussion	France	1.7	0.7	3.1	4.2
	Spain	6.3	2.1	2.2	3.9
	Belgium	10.4	1.7	9.1	16.8
News on TV	France	6.6	0.9	0.9	4.7
	Spain	3.8	1.1	4.7	8.4
	Belgium	3.8	0.7	2.3	3.7
News daily papers	France	7.8	0.6	2.6	10.4
	Spain	8.1	3.4	5.2	10.2
	Belgium	6.9	1.4	6.9	6.4
News on radio	France	13.8	0.9	6.9	7.6
	Spain	11.2	1.0	6.7	7.8
	Belgium	13.3	1.0	2.7	10.5
Interest in European politics	France	10.5	1.0	7.4	2.9
	Spain	13.8	1.6	8.6	6.2
	Belgium	9.3	2.1	2.4	7.3
Level of EU informedness	France	8.4	0.7	6.1	5.7
	Spain	11.1	1.4	12.0	6.1
	Belgium	11.0	0.7	15.1	9.3
Membership in EU	France	6.7	0.5	6.0	12.3
	Spain	7.0	0.8	6.3	9.9
	Belgium	16.6	2.0	7.0	20.2
Benefit from EU membership	France	7.7	0.6	1.3	7.1
	Spain	14.0	0.7	4.3	13.6
	Belgium	15.1	2.4	3.9	10.9
Colour TV	France	4.1	0	2.6	6.7
	Spain	0.7	0.5	1.0	2.2
	Belgium	2.5	0.2	1.7	4.4
PC	France	0.2	0.5	5.3	5.6
	Spain	7.4	1.5	4.2	1.7
	Belgium	8.4	2.4	3.1	2.9
Two or more cars	France	4.7	1.3	2.8	0.6
	Spain	10.6	1.9	1.5	7.2
	Belgium	9.3	2.2	1.2	5.9
Second home	France	0.8	0.2	2.0	1.4
	Spain	4.9	2.0	2.6	0.3
	Belgium	1.0	0.4	0	0.6
Mean	France	7.1	0.6	4.3	6.9
	Spain	9.4	1.4	5.3	7.4
	Belgium	8.2	1.4	6.2	8.9
Average across three countries		8.2	1.1	5.2	7.7

These effects differ from question group to question group. This is not surprising because the strength of the effects is always dependent on the strength of the relationship between the error source and the substantive type of variable, and this differs from topic to topic. Nevertheless, it is clear that the coverage error is the smallest problem and that the two other factors can produce quite large differences between studies done with different modes of data collection or by different organisations. In general the effects are so large that without correction the results cannot be compared. Therefore, more attention will be devoted to mode effects as the second largest source of differences. Unfortunately nothing more can be said about the organisational differences than what has already been remarked in chapter 2. As a consequence, this chapter will now concentrate on mode effects and on possible correction for differences between studies in general.

11.3.2 The mode effects

Given the considerable contribution of the pure mode effects to the total differences between the results obtained with the two data collection methods, a more detailed analysis of the pure mode effects was conducted on the basis of the panel study.

The mode effects are different for different types of questions, suggesting a separate look at open-ended questions, simple closed questions and complex closed questions.

Starting with the open-ended question, the study reported in chapter 7 concentrated on an agenda question asking for the two most important problems for the own country and for Europe. The study evaluated three aspects: the mode effect on nonresponse, frequency with which the different problems were mentioned, and the amount of information provided by the respondents.

First, the amount of nonresponses in the two survey modes was analysed. No substantial differences in the data were found. The second step dealt with the content (or quality) of the answers. Here, much to the authors surprise the expectations were reversed: more diverse answers were obtained in telephone surveys than in face to face interviews. Large differences were found for the issues of employment and health. The similarity in results for the two telephone surveys seems to suggest that there are indications of pure mode effects in these cases. These differences did not change the ordering of the importance of the issues but could produce such a change quite well if more subcategories are used for the unemployment problem.

Finally, the talkativeness (and the possible obstructions for talkativeness) in respondent behaviour was considered. Clear differences between the two polling firms were found. Due to the lack of data it could not be determined whether the differences were due to agency effects or mode effects.

Among the complex close-ended questions, the left-right orientation has been highlighted in chapter 8. Normally in face to face research a show card is used which presents the 10 categories ranging from the extreme left to the extreme right. Some researchers has suggested that in telephone interviewing this approach is not feasible, given the complexity of the question and the large number of response categories. This has lead to the idea to apply a two-

step procedure. Fortunately, in this study a comparison could be made between the standard question and the two step procedure in face to face interviews and between the standard 10 point scale in face to face and the 10 point scale in telephone interviewing. It was found that the change of format led to much larger differences in the response distributions in the different countries than the use of the 10 point scale in the different modes of data collection. Thus, strong evidence suggests that the two-step procedure for the left-right scale is not a good telephone alternative for the standard 10 point scale in face to face interviews. The results obtained with the two-step-version will not be comparable. On the other hand, the two 10 point scales remained more comparable for the two different modes of data collection, although also in that case significant differences between the two could be detected. As a consequence, corrections for these differences are required in order to compare the results obtained with the same scale in different data collection modes.

With respect to the standard Eurobarometer questions, mode effects occurred for some questions and not for others. In chapter 6, for the evaluation of the EU membership significant differences between face to face and telephone interviewing have been found, as was true for the satisfaction questions and for the persuade questions.

Using the same approach, in chapter 9, a larger set of close-ended questions was scrutinised whether there was a mode effect and also whether there was a difference in response probabilities in the different countries. This last point is interesting because comparison of responses across countries is only possible if in the different countries response probabilities, i.e. the probability of a specific answer given the (latent) opinion of a person, are the same. For example, regarding the "benefit question" the results are only comparable if people in different countries believing that their country has benefited from EU membership have an equal probability to also say that their country has benefited. If these response probabilities vary, the difference in responses does not result from a difference in opinion but from a difference in response probabilities. In table 11.2, the results of the respective test in this study is presented once more.

Starting with the ***media involvement*** questions, they have not been affected by the mode of data collection. The categories are relatively detailed and require separate estimates of frequencies, but apparently the mode of data collection has no effect. It is also important to note that there are no differences in response probabilities across countries. So the response categories for these questions can be used for comparison across modes and across countries.

The second set of questions concerns ***political involvement***. One of the two questions has the same response probabilities for the different modes and different countries, while the other produces differences between modes and across countries. The reason for this difference is that the question on "political discussion" entails the categories frequently, occasionally, never, and a DK/No answer, and the question "persuade" the often, from time to time, rarely, never, and a DK/No answer, that is one additional category. Comparing the meaning of the categories in the two questions just on face value, one could conclude that "rarely" is an unnecessary extra category. In fact, it turns out that it is precisely this category which causes the differences between the modes and the countries. Consequently, if the categories "rarely" and "never" are collapsed, the problems might disappear. After a test, it turned out indeed that

the model with equal probabilities across modes and across countries now fitted the data. This suggests that for purposes of comparison the categories “rarely” and “never” should be combined. To leave the category “rarely” completely out in the data collection is, of course, another viable option.

The third set of questions concerns ***satisfaction***. These two questions have unequal response probabilities across modes and countries. Looking at the response categories, however, the problems do not come as a surprise. The response categories for both questions are: very satisfied, fairly satisfied, not very satisfied, not at all satisfied, and a DK/No answer. The problems are that it can be confusing for translators and respondents how to interpret especially the labels “fairly satisfied” and “not very satisfied”. In fact, logically one could argue that after the “very satisfied” category “not very satisfied” contains all other possible answers and that therefore it is not clear when to use the category of “fairly satisfied”. In the latent class analyses it was indeed found that people in category 3 on the latent variable had a different probability for answering fairly satisfied and not very satisfied in the different modes. But this problem may also carry over into the translation of these categories into different languages.

Checking this hypothesis, differences in the translations in the different languages were indeed discovered. In French and Dutch, the translation of the labels were as follows: very satisfied, rather satisfied, rather dissatisfied, not at all satisfied. This is quite different from very satisfied, fairly satisfied, not very satisfied, not at all satisfied. If such differences exist in the translation between countries, it cannot come as a surprise that also differences in the reactions of the respondents across countries are found. The differences across modes must also have to do with the above mentioned problematic categories which may work differently in face to face than in telephone research.

Interesting in this respect are the findings for the ***involvement in the EU***. The question on knowledge uses the same category system as the satisfaction questions: very well informed, quite well informed, not very well informed, not at all well informed, and DK/No answer. So, if the above interpretation is correct, this question should have the same problems as the satisfaction question. In table 11.2 one can see that this is indeed the case, strengthening the argument given before.

The question on “interest in EU matters” uses different labels: a great deal, to some extent, not much, not at all, and a DK/No answer, but the problem is comparable. The term “not much” is a negation of “much”. So if one is less than “much interested” in the EU, one could choose “not much”. But then the position of the category “to some extent” is again not clear. It can be regarded as a part of the category “not much”, but that would lead to confusion. Thus, the same problems as for the other questions were expected and were indeed found, as can be seen in table 11.2. Checking the translations in the different countries, it was found that also in this case the translation of the labels is not equivalent.

Finally, seven ***opinion*** questions, all with the same format “pro, against and no opinion”, were analyzed. According to table 11.2, the questions concerning the introduction of the European Monetary Union and a common defense policy for the EU did not indicate any mode effect and differences across countries. On the other hand, all questions concerning the

elections showed a cross-cultural mode effect. In the telephone interview people with a score of “DK/No answer” on the latent variable have a higher probability to say “pro” than in the face to face interview. Besides that, the question concerning the division of tasks between local, national and EU government produced the same result.

It is difficult to explain these effects. It cannot be a general acquiescence bias (Schuman and Presser, 1981) because then it should occur for all questions. It is also not an effect of the topic because then one would have to find a different explanation for the last question. An interpretation as a learning effect is also questionable because then one would also expect this outcome for all topics and not only for a limited number. Besides that, there is no obvious reason why all people should learn that they have to respond “yes” to this question instead of “no”. So for the time being one has to accept these findings and will have to wait for further research to clarify the matter.

11.3.3 Adjustment for mode differences

Finally, an effort has been made in chapter 10 to see if it would be possible to adjust the results obtained with face to face interviewing to the results obtained with telephone interviewing. This activity makes sense because it is very inconvenient for users that the two data collection modes produce different results for the same questions. This could lead to a lot of confusion if these results would be reported without further comment. Imagine that the standard Eurobarometer (face to face) would present a positive opinion of the public in one month, and in the next month the telephone tracking study would present a much more negative opinion. Is there reason for alarm or not politicians and journalists might ask. In such a situation one has to be able to correct the tracking study data for the mode effects in order to check whether there are indeed significant differences to the standard Eurobarometer once this correction has been performed.

Chapter 10 suggests that a procedure can indeed be found which transforms the results from the telephone interviews to the results from the face to face interview and vice versa. However, this procedure needs to be specific for the response-nonresponse structures of those companies conducting the studies. This means that the numeric solution obtained in chapter 10 cannot be generalised to data collections which will occur in the future. Rather, they must be estimated again for new data because different companies collect these data. Furthermore, this chapter has suggested that the adjustments are specific for each question because the effects of coverage errors, nonresponse and mode effects are different for each question. The most important finding, however, is that such adjustments are possible. How stable the results are across time is an interesting issue for further research.

Table 11.2 The evaluation of differences between modes and countries for EB41.Panel

Variables	Equality of the parameters across countries		not across countries	
	across modes (model 1)	not across modes (model 2)	across modes (model 3)	not across modes (model 4)
Media involvement				
Radio	+			
Newspaper	+			
TV	+			
Political involvement				
Political discussion	+			
Persuade others	-	-	-	+
Satisfaction				
Life in general	-	-	-	+
Democracy in country	-	-	-	+
Involvement in EU				
Interest -	-	-	+	
Knowledge	-	-	-	+
Opinion on EU membership				
Benefit for country from EU membership	+			
Evaluation of membership for country	-	+		
Opinion on EU policies				
European Monetary Union	+			
EU defense	+			
Participation local elections	-	+		
Participation EU elections	-	+		
Candidacy local elections	-	+		
Candidacy EU elections	-	+		
Division of tasks between various levels of government	-	+		

11.4 Conclusion

In this study the comparability of the responses in surveys across modes of data collection and across countries was studied, and considerable differences were found. In this last section some consequences of this finding for survey research will be drawn, and some scientific issues will be discussed.

11.4.1 Practical consequences

The results clearly indicate that quite large differences can be expected for many questions of the Eurobarometer surveys if one compares the results of the standard face to face Eurobarometer and the telephone-based tracking study.

It has been shown that there are at least three factors which can explain why these differences occur. For the Eurobarometers where two different companies do the face to face study and the telephone study, it is hard to suggest ways to reduce these errors by adjustment of the existing procedures. Partially the differences are due to conditions which exist in telephone interviewing (many call-backs) but which do not exist in face to face interviewing, and vice versa (show cards).

Some of the mode effects can be avoided. It was found that for four questions the mode effects had to do with the formulation of the response categories. For these four questions the category labels were overlapping which led to confusion in the translations and to differences between modes. With adjustment of the categorisations these problems can be avoided in the future. For example, for the satisfaction question in some countries the labels very satisfied, fairly satisfied, not very satisfied, not at all satisfied should be substituted by the labels very satisfied, rather satisfied, rather dissatisfied, very dissatisfied.

Quite likely this categorisation will lead to less problems if used in all countries. The substantive disadvantage of this categorisation is that since usually more people in Europe are satisfied than dissatisfied, this scale is de facto used only as a two-point scale.

An alternative categorisation might be very satisfied, rather satisfied, little satisfied, not satisfied. Both formulations will probably be comparable across modes and countries, but this has to be tested. This categorisation will also lead to less confusion accross modes. For the two questions about involvement in the EU similar labels should be used to avoid problems.

It was also found that simple weighting procedures cannot be used to correct for the differences between the two types of studies. However, for all questions the correction procedure can be used which has been discussed in chapter 10 to make the results of the standard Eurobarometer as comparable as possible to a telephone study. For this purpose the pure mode effects and the coverage error as estimated in this study can be used while the fieldwork effects have to be calculated again because fieldwork organisations will probably change across time. If the connected “nonresponse’ difference has been assessed, the correction can be done for all questions in the same way as was demonstrated in chapter 10. If the procedures used by field organisations are not changed, these new estimates should be expected to remain the same, and the corrections can be used until a change in the procedure is introduced. At that moment again new estimates have to be obtained.

11.4.2 Scientific issues

Table 11.1 shows that large differences are found across modes for different variables. While the methodological literature gives most attention to coverage errors, it turned out that the coverage errors are relatively small compared with the other two types of errors. Especially

the effects of the procedures used by the different research organisations are relatively large. Unfortunately, this effect because of the many contributing factors cannot be decomposed into the different possible causes in the present study.

Table 11.2 has shown that for 11 out of the 18 questions the responses are effected by the pure mode of data collection when estimated on the basis of the panel data. This finding which is in agreement with the results in chapter 5 using a different approach, is also contradicting some of the standard literature on mode effects (Groves and Kahn, 1979; de Leeuw and van der Zouwen, 1988; de Leeuw, 1990) but agrees with other studies which found considerable effects (Silberstein et al., 1989; Kalfs, 1994; Scherpenzeel and Saris, 1997). This last study is the most comparable one to the one discussed in this book since there also panel data have been used. In panel studies the confounding factors can be better controlled although there one faces the additional problem of memory effect. In that study also mode effects were discovered, for example for the satisfaction variables. In a meta analysis of similar studies (Scherpenzeel 1995) in different countries also country-specific effects have been found like the ones reported in table 11.2. Such country-specific differences occur much less frequently than mode effects. For only five questions the response probabilities across countries were different. In four questions the category labels were overlapping which led to confusion in the translations and in the responses. These confusing category labels apparently are the major reason for the cross-cultural differences obtained. It would be desirable to pursue this idea in further research.

In this context the importance of the equality of the response probabilities has been emphasised because cross-cultural comparison requires these equalities. If they have not been found, then observed differences can be explained by the differences between the response probabilities instead of the differences in opinions. The analysis in chapters 6 and 9 has shown that the test of such equalities is relatively easily done using the latent class model and the program LEM (Vermunt, 1996).

In sum, this book has served three major purposes. For one, the methodological study reported here has indicated that it is worthwhile, consequential and possible for survey research to systematically study effects of coverage, interview modes and nonresponse on the answers of those questioned. There is no excuse for methodological naiveté in the matter. Secondly, it has pointed to the need to consistently continue to pursue these effects in ongoing everyday research in order to present methodologically enlightened findings to the public and to the clients. Thirdly, the study has also indicated where additional methodological research is necessary. It is hoped that this book contributes to all three of those prerogatives.